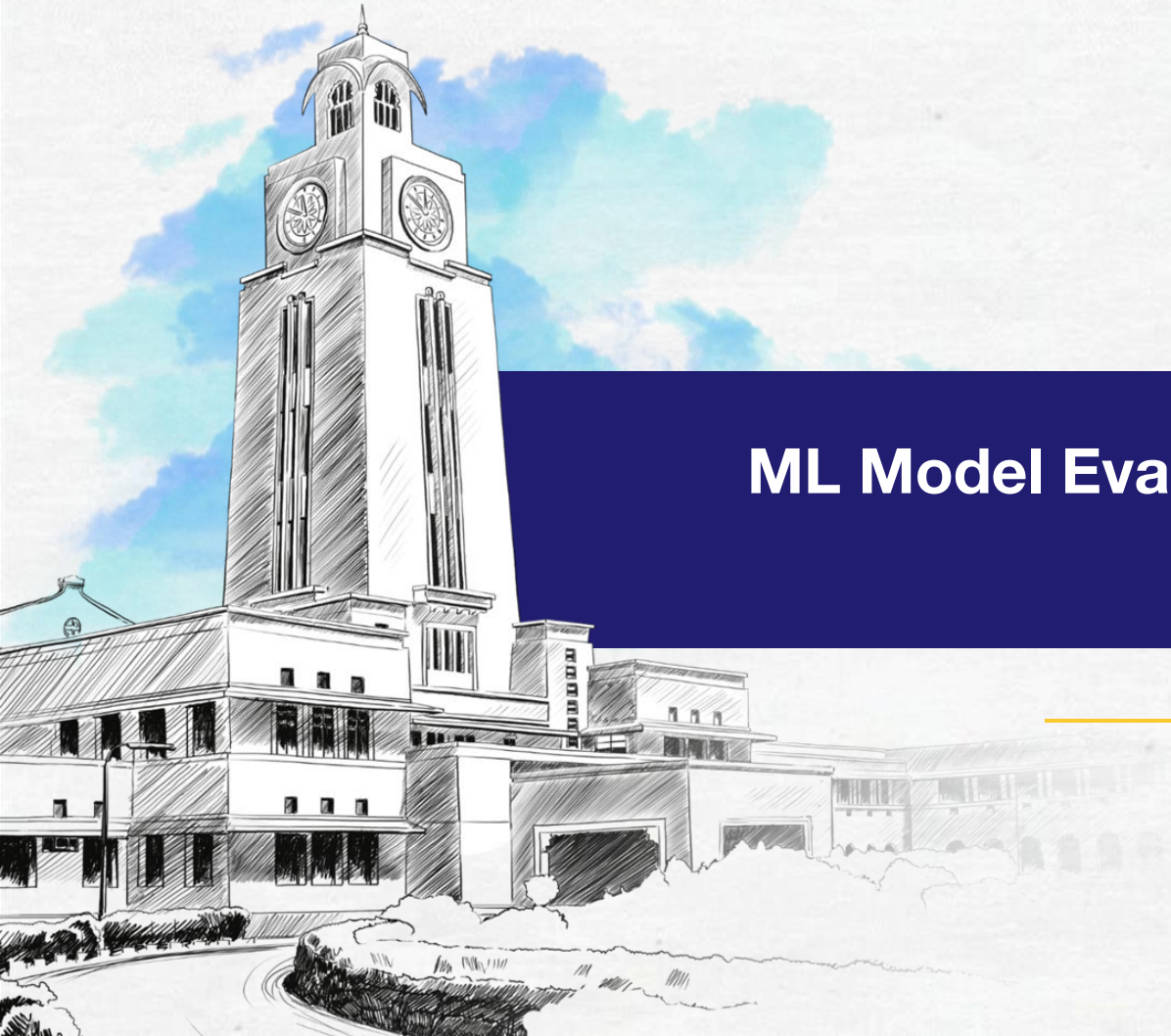


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

ML Model Evaluation & Introduction to FACCT ML

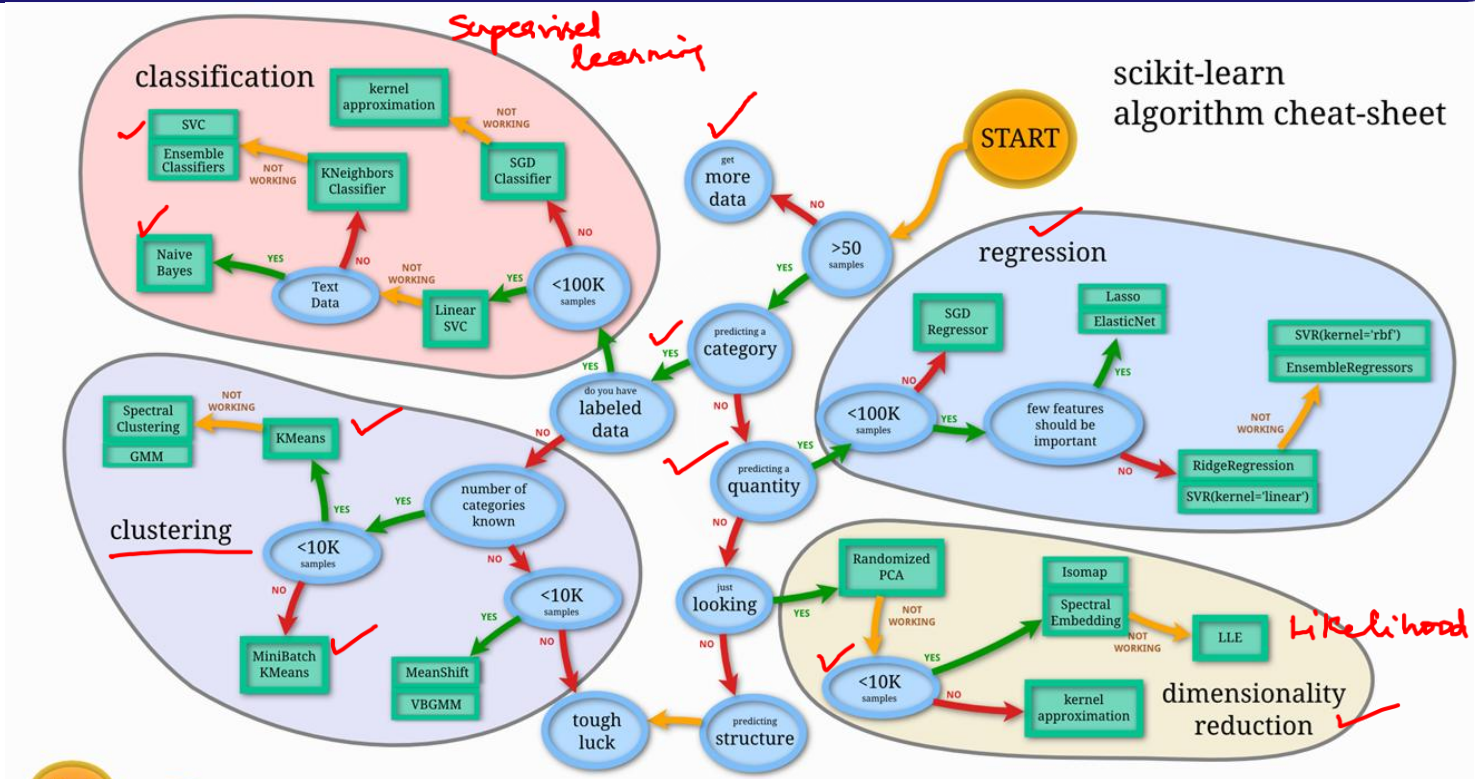
Dr. A.R.Kiruthiga

Guest Faculty
BITS Pilani

Agenda

- Comparing ML Classifiers (Was already covered in respective modules. Confusion matrix-based metrics , RMSE, R2, AUC-ROC, Cross Validation)
- Emerging Requirements
  - Bias and Fairness
  - Interpretability

Model Guide



Supervised learning

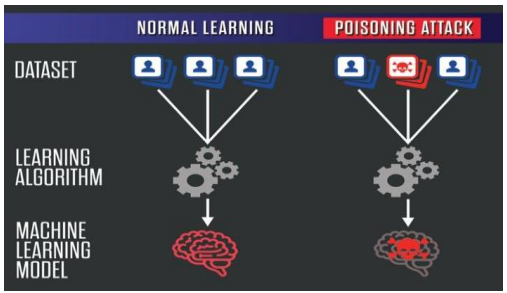
Likelihood Estimation

Back

Source Credit : https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

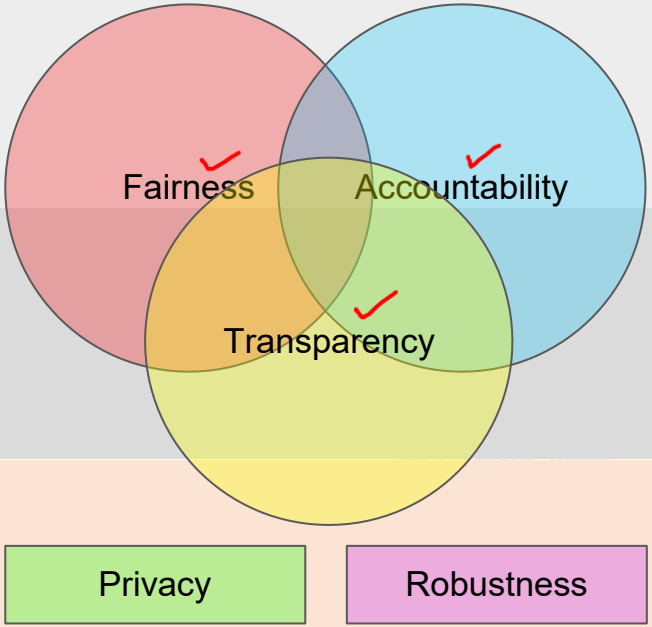
Emerging Requirements

- **Fairness**
  - What to do to ensure gender and ethnic fairness in ML models?
- **Accountability**
  - Who takes the responsibilities for failed ML models?
- **Transparency**
  - What to do to make ML models transparent and comply with regulations?
- **Privacy Issues**
  - How to protect user privacies when exposing data to ML models?
- **Security Issues**
  - How do we defend ML models against data poisoning?



FAccT Overview (More on this is a part of next semester elective)

Different academic areas contributing to FAcCT:



Psychology
Social Science
Public Policy

Statistics
Theory

Machine
Learning

Top Layer: Human & Social Perspectives

Psychology

Social Science

Public Policy

These fields study:

human behavior,

ethics,

societal impact,

regulations.

Middle Layer: Mathematical Foundations

Statistics

Theory

These provide:

fairness metrics,

probabilistic reasoning,

formal guarantees.

Bottom Layer: Machine Learning

This is the technical implementation layer where:

algorithms are built,

models are trained,

fairness/privacy techniques are applied.

This emphasizes that trustworthy AI is **not only a machine learning problem**.
Building responsible AI requires:

- technical methods,
- mathematical theory,
- ethical thinking,
- legal and policy understanding,
- social awareness.

FAcCT is therefore a **multidisciplinary field** focused on making AI systems:
fair,
explainable,
accountable,
privacy-preserving,
and robust.

Real World Example

Amazon's Secret AI Hiring Tool Reportedly 'Penalized' Resumes With the Word 'Women's'

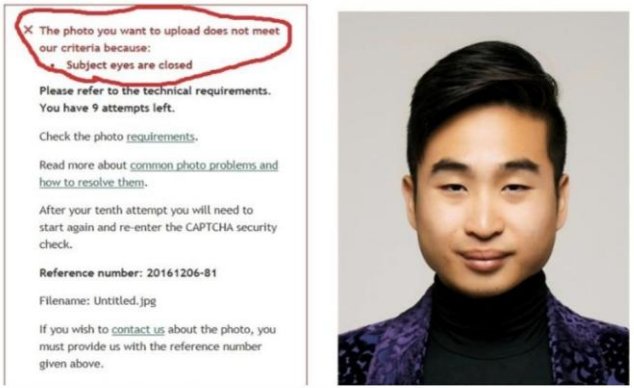
Rhett Jones
10/10/18 10:32AM • Filed to: ALGORITHMS

79 3

Photo: Getty

New Zealand passport robot thinks this Asian man's eyes are closed

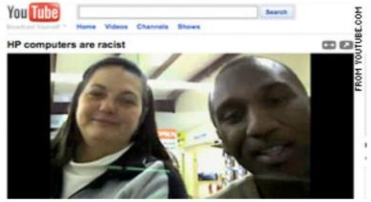
By James Griffiths, CNN
Updated 1:46 AM ET, Fri December 9, 2016


The photo you want to upload does not meet our criteria because:
Subject eyes are closed
Please refer to the technical requirements.
You have 9 attempts left.
Check the photo requirements.
Read more about common photo problems and how to resolve them.
After your tenth attempt you will need to start again and re-enter the CAPTCHA security check.
Reference number: 20161206-81
Filename: Untitled.jpg
If you wish to contact us about the photo, you must provide us with the reference number given above.

New Zealand's online passport application system couldn't recognize Richard Lee's open eyes.

HP looking into claim webcams can't see black people

By Mallory Simon, CNN
December 23, 2009 7:25 p.m. EST



an HP webcam
tware.

(CNN) -- Can Hewlett-Packard's motion-tracking webcams see black people? It's a question posed on a now-viral YouTube video and the company says it's looking into it.

In the video, two co-workers take turns in front of the camera -- the webcam appears to follow Wanda Zamen as she sways in front of the screen and stays still as Desi Cryer moves about.

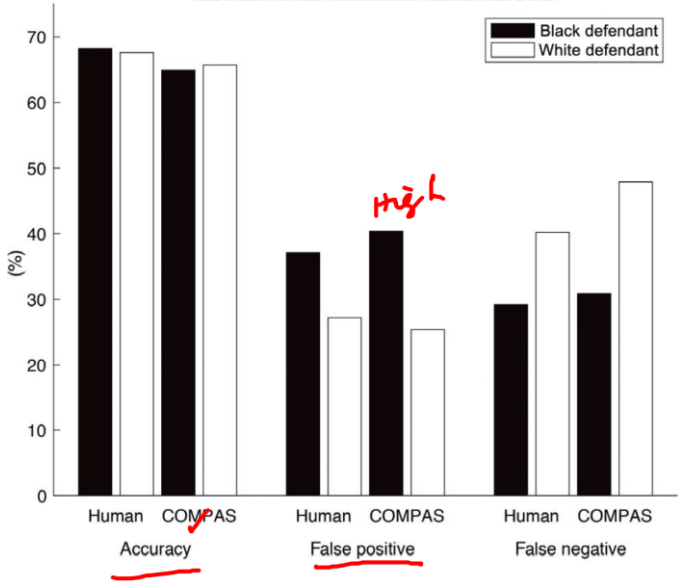
HP acknowledged in a statement e-mailed to at the cameras may have issues with contrast recognition in lighting situations. The webcams, built into HP's new ers, are supposed to keep people's faces and bodies in on and centered on the screen as they move.

eo went viral over the weekend, garnering more than 400,000 e page views and a slew of comments on Twitter.

Algorithmic Bias

Commercial risk assessment software known as COMPAS used in the criminal justice system.

- Assess more than 1 million offenders since 2000
- Predicts a defendant's risk of committing a misdemeanor or felony
- 137 features



Dressel et al, 2018

- Can an AI system make unfair predictions for different racial groups?
- Researchers found that COMPAS showed evidence of racial bias.

False Positive Rate

A **false positive** means:

The system predicts someone is high risk, but they actually do NOT reoffend.

The graph shows:

•Black defendants had **higher false positive rates**.

•COMPAS more often incorrectly labeled Black defendants as dangerous/high-risk.

This is considered unfair because innocent people may receive:

•harsher bail,

•longer sentences,

•stricter supervision.

Even though overall accuracy was similar, the **errors were distributed unequally across racial groups**.

Why This Matters

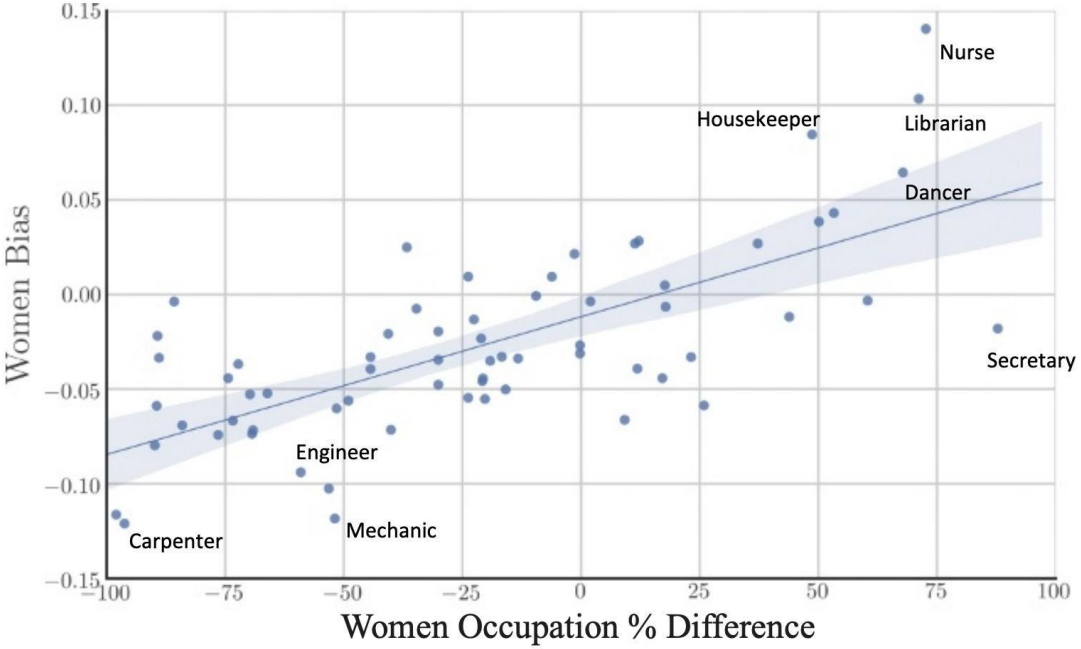
The COMPAS case became a famous example in AI ethics because it showed that:

- algorithms can inherit societal biases,
- fairness is difficult to define mathematically,
- automated decisions can affect real lives.

It raised concerns about:

- discrimination,
- transparency,
- accountability,
- fairness in criminal justice AI.

Bias in Historical Data – Negative Legacy



Gard et al, 2018

Fairness

https://www.borealisai.com/research-blogs/tutorial1-bias-and-fairness-ai/

- What is Fairness?
  - The absence of bias towards an individual or a group (Mehrabi et al, 2019)
- Can ML Models Discriminate?
  - Aren't machines just follow human's instructions?
  - ML models approximate patterns in the data
  - Learns/Amplifies biases at the same time

Yes, Machine Learning (ML) models can discriminate if they learn biased patterns from data or are designed without fairness considerations.

Why discrimination happens?

ML models learn from historical data. If the data contains human or societal biases, the model may reproduce or even amplify them.

How to reduce discrimination

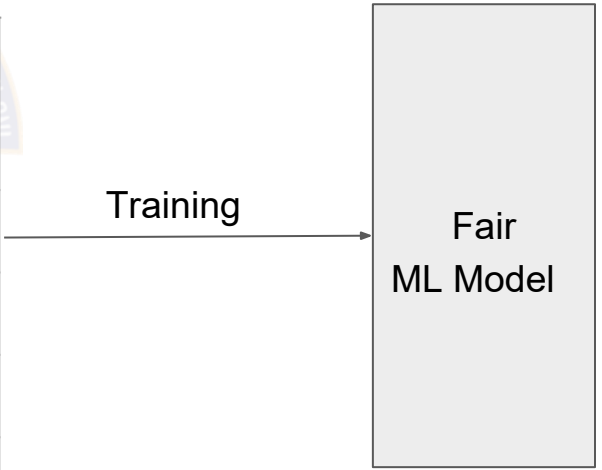
- Use diverse and representative datasets.
- Detect and measure bias using fairness metrics.
- Apply interpretable and transparent models.
- Audit models regularly.
- Include human oversight and ethical review.
- Remove or control sensitive and proxy features.
- ML interpretability and fairness techniques are important because they help identify when and why discrimination occurs.

Fairness Through Unawareness

A ML Algorithm Achieves Fair Through Unawareness If

- None of the sensitive features are directly used in the model

~~Race and Ethnicity~~	Skills	Years of Exp	Hired?
Hispanic	Javascript	1	no
Hispanic	C++	5	yes
White	Java	2	yes
White	C++	3	yes



- Sensitive Features May Still Be Used
  - Inferred from indirect evidence

• Protected

• Inferred

Race and Ethnicity	Skills	Years of Exp	Often Goes to Mexican Markets	Hiring Decision
Hispanic	Javascript	1	yes	no
Hispanic	C++	5	yes	yes
White	Java	2	no	yes
White	C++	3	no	yes

Training → Discriminatory ML Model

Need for ML Interpretability

Our society has been shifted to rely on AI more than ever

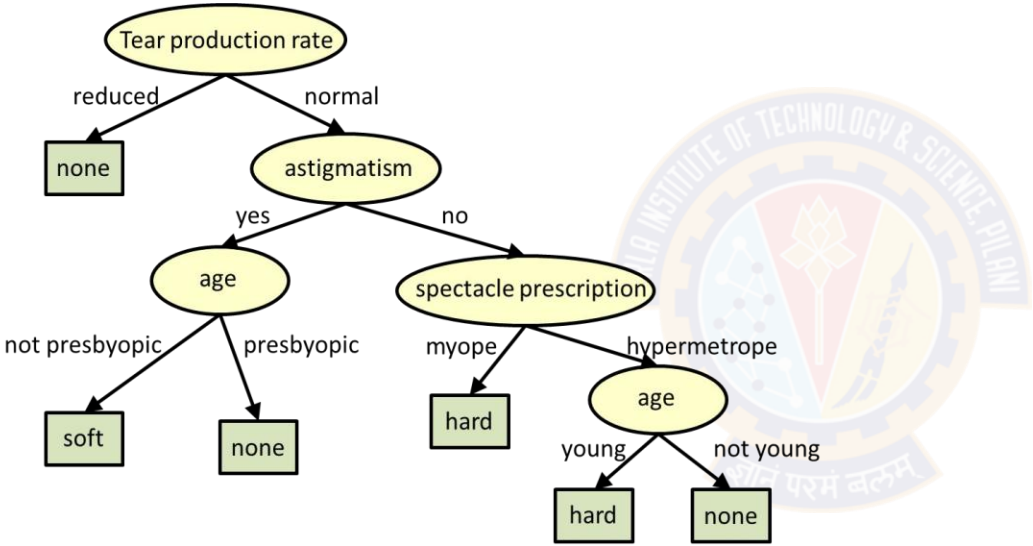
- autonomous vehicles
- Security
- Finance
- many others

Who will benefit from ML Interpretability?

- **End Users**
  - Enhances trust in AI systems.
  - Helps users understand the consequences of model decisions (e.g., impacts on privacy, fairness, or credit approval).
  - Increases transparency and user confidence.
- **Regulatory Agencies**
  - Supports compliance with laws and regulations.
  - Enables auditing of AI systems.
  - Improves accountability and governance of automated decisions.
- **Model Designers / Data Scientists**
  - Helps diagnose and debug model performance issues.
  - Explains why a model makes certain predictions.
  - Identifies bias, errors, overfitting, or unintended behavior.
  - Improves model reliability and robustness.

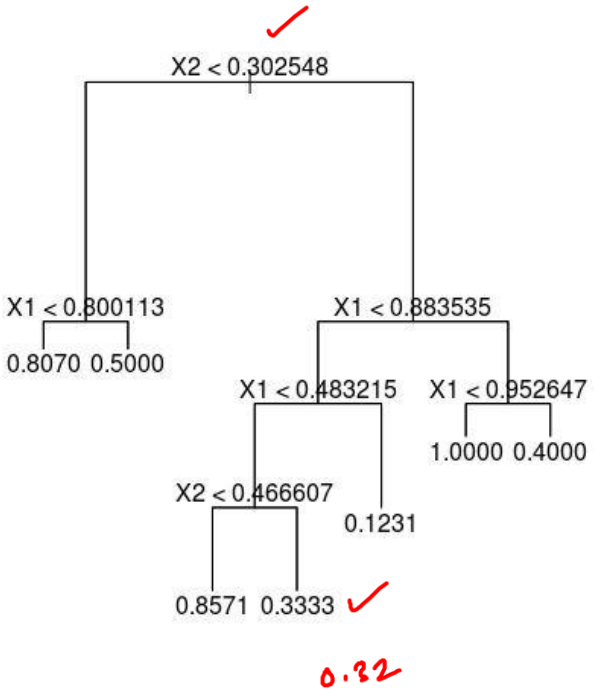
Interpretability in Decision Tree

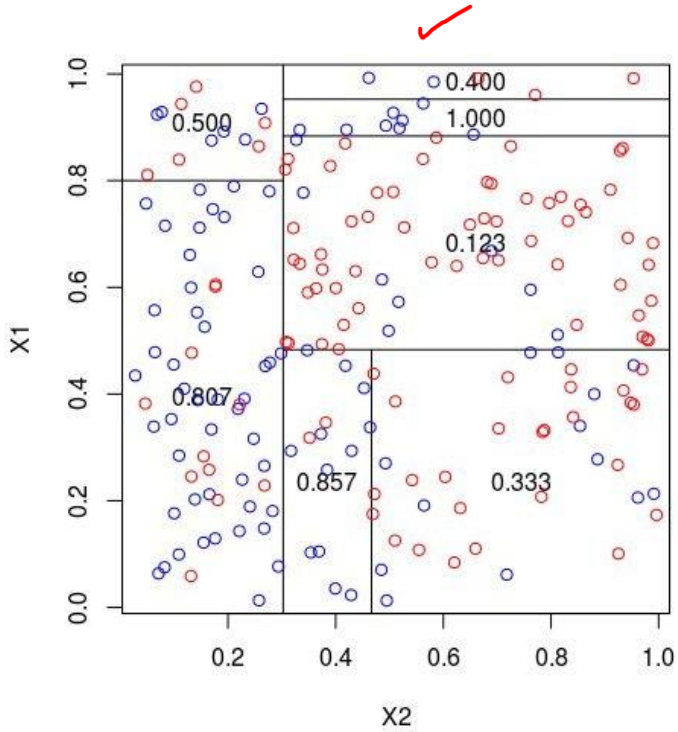
ML interpretability allows one to examine model's basis in its decision making process



An interpretable tree model to find out the kind of contact lens a person may wear

Interpretability in Decision Trees



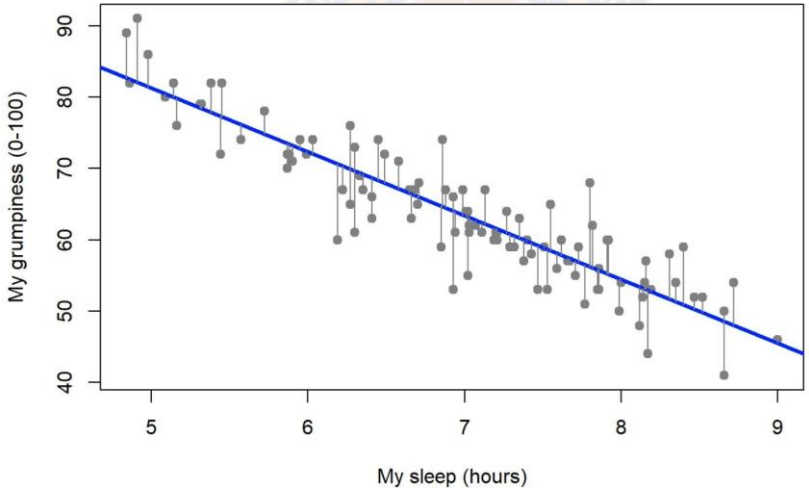


Interpretability in Linear Regression

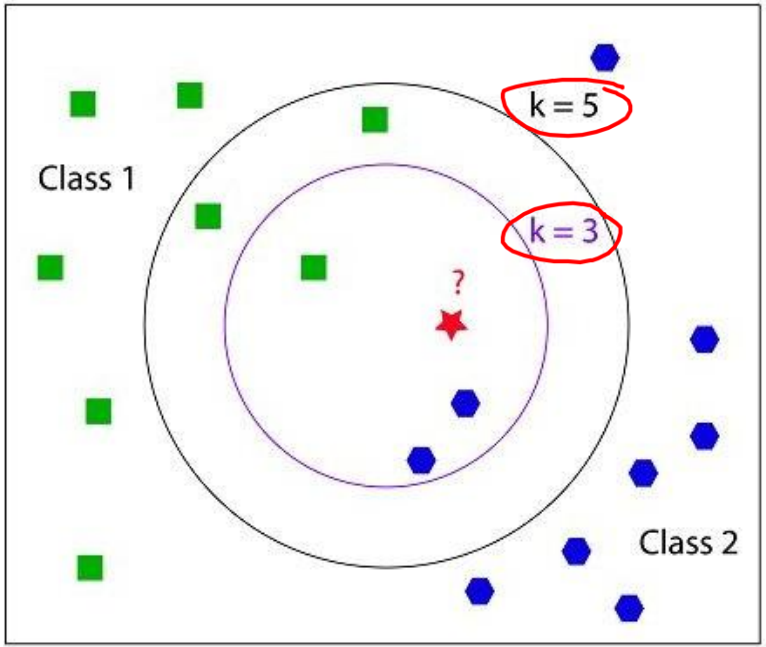
$a_0 + a_1 x_1 + a_2 x_2 + \ldots$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

interpretable components

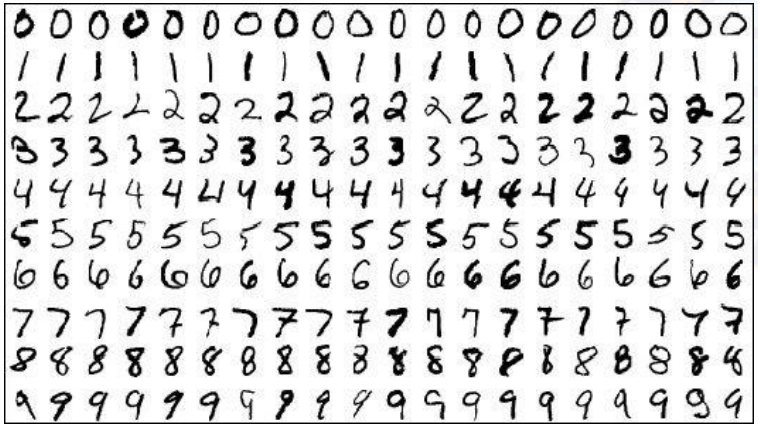


Interpretability in K-Nearest Neighbors

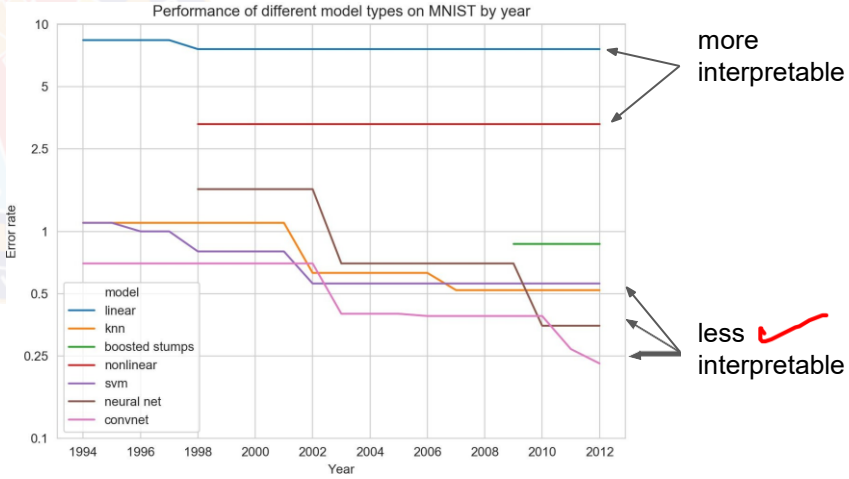


Interpretability and Performance Trade-offs

- Highly performed models tend to be less interpretable.
- Can powerful models with complex structures be interpretable at the same time?

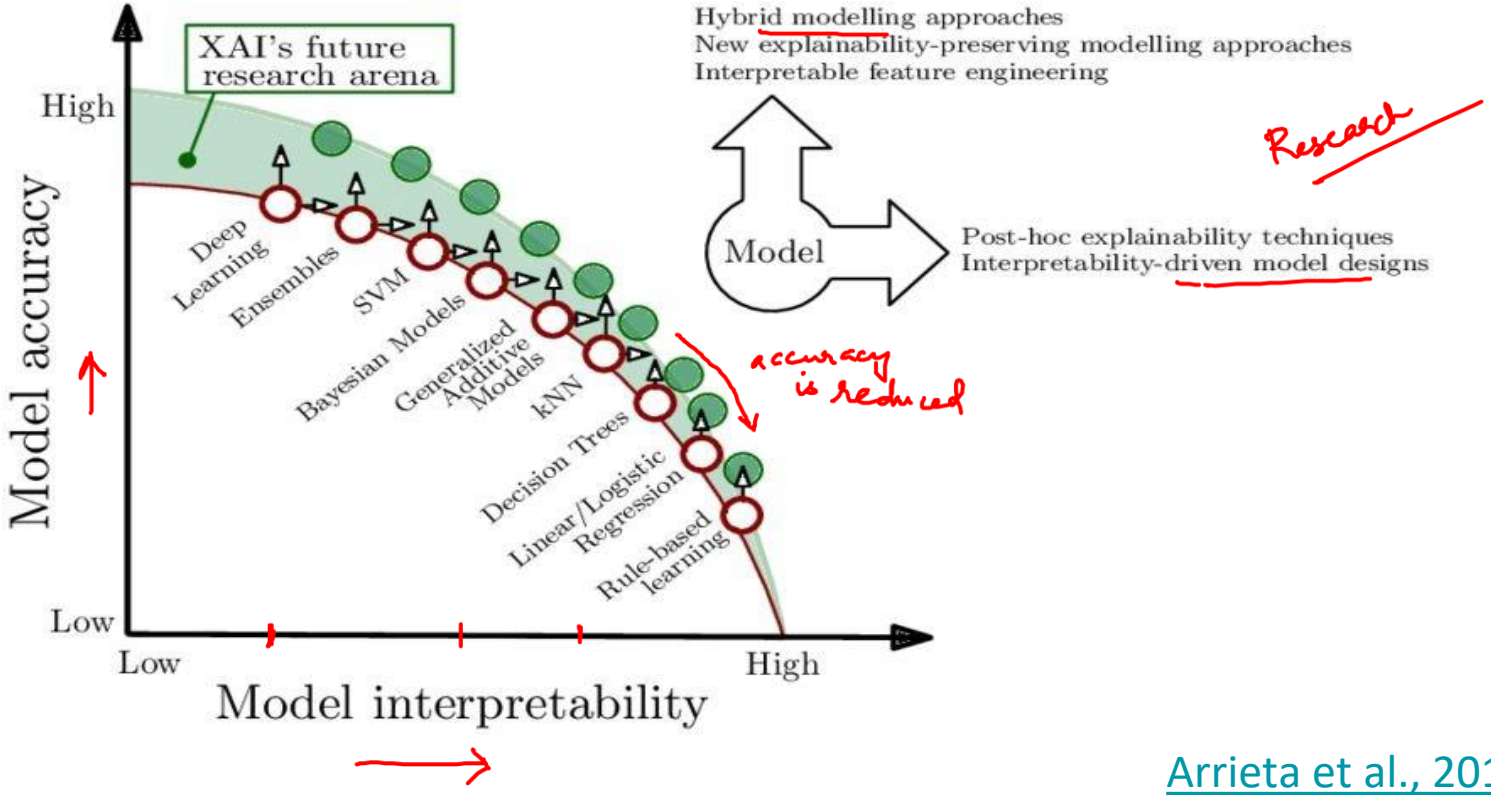


MNIST Dataset



http://yann.lecun.com/exdb/mnist/
https://soph.info/2018/11/08/mnist-history/

Interpretability and Performance Trade-offs



Arrieta et al., 2019